CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 13**

Risk bounds, concentration inequalities

Lecturer: Peter L. Bartlett *Scribe: Zile Wei*

1 Risk bounds

We continue on examining the relationship between the empirical risk and the true risk. So far, we have looked at algorithms that optimize over a set of training samples,

$$\min_{f \in \mathcal{F}} \hat{R}(f) = \hat{E}l(y, f(x)) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i)) \tag{1}$$

How does the minimizer f^* over empirical risk tell us about the true risk $R(f) = El(y, f(x))$? Particularly, we expect that (1) will converge to true risk as $n \to +\infty$, under some proper conditions. But we need to be careful about the richness of function class $\mathcal{F}$ and how the difference between empirical risk and true risk depends on the number of training samples, n.

For example, let $\hat{E}_\phi(f) = \hat{E}\phi(yf(x))$ and $R_\phi(f) = E\phi(Yf(X))$. For a single f, by the law of large numbers, the empirical average will converge to the expectation, or true risk. $Z_i = l(y_i, f(x_i))$, $E|Z_i| < \infty \Rightarrow \frac{1}{n}\sum Z_i \xrightarrow{a.s.} EZ_i$

But if $\mathcal{F}$ is too large, there are some cases that the empirical risk will not converge to the true risk. That is

$$\sup_{f \in F} |\frac{1}{n}\sum l(y_i, f(x_i)) - El(Y, f(X))| not \to 0$$

e.g. $\mathcal{F}$ is all measurable functions. $f : X \to \{\pm 1\}$. Suppose $P(Y = 1|X = x) = 1$ a.s. We can choose $f(x)$ as $f(x_1) = \cdots = f(x_n) = 1$ and $f(x) = 0$ otherwise. Hence $El(Y, f(X)) = 1$, but $\hat{E}l(Y, f(X)) = 0$.

If $\mathcal{F}$ is finite, the empirical risk will converge to the true risk. Since for every single f in $\mathcal{F}$, we can say that its empirical risk converges to true risk. When we minimize over finite finite function class $\mathcal{F}$, the convergence is guaranteed.

2 Concentration inequality

Concentration inequality will tell us how rapidly the empirical risk converges to true risk. We begin with some loose upper bounds.

Theorem 2.1. Markov inequality

For any almost surely nonnegative random variable X, and $t > 0$.

$$P(X \geq t) \leq \frac{EX}{t}$$

Proof.

$$EX \geq E(X\mathbb{1}[X \geq t]) \geq tP(x \geq t)$$

Therefore

$$P(X \geq t) \leq \frac{EX}{t}$$

$\square$

Theorem 2.2. Chebyshev inequality

For any random variable X and $t > 0$,

$$P(|X - EX| > t) \leq \frac{Var(X)}{t^2}$$

Proof. Directly apply Markov inequality on random variable $|X - EX|^2$. $\square$

For example, binary random variable $Z_i \in \{0,1\}$, $p = P(Z_i = 1)$, $Z_1, \cdots, Z_n$ are i.i.d. random variables. $S_n = \sum_{i=1}^n Z_i$, $\sigma^2 = Var(Z_i) = p(1-p)$, $Var(S_n) = n\sigma^2$. From Chebyshev inequality, choosing $t = n\epsilon$, we have

$$P\Big(\Big|\frac{S_n}{n} - \frac{ES_n}{n}\Big| \geq \epsilon\Big) \leq \frac{\sigma^2}{n\epsilon^2} \tag{2}$$

But actually, the upper bound on (2) is quite weak. From central limit theorem,

$$P\Big\{\sqrt{\frac{n}{\sigma^2}}\Big(\frac{S_n}{n} - p\Big) \geq y\Big\} \to 1 - \Phi(y) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-y^2/2}}{y}$$

So

$$P\Big\{\frac{S_n}{n} - p \geq \epsilon\Big\} \approx e^{-n\epsilon^2/(2p(1-p))} \tag{3}$$

(3) decreases exponentially instead of polynomially in (2).

Hoeffding's inequality Concentration on the sum of independent random variables

Suppose random variables, $X_1, \cdots, X_n$, are independent, $S_n = \sum_{i=1}^n X_i$.

For

$$\begin{aligned}
P\{S_n - ES_n \geq t\} =& P\{\exp\big(s(S_n - ES_n)\big) \geq e^e t\} \\
\leq& e^{-st}E[\exp(s(S_n - ES_n))] \\
=& e^{-st}E\exp\Big(s\sum_{i=1}^n (X_i - EX_i)\Big) \\
=& e^{-st}\prod_{i=1}^n Ee^{s(X_i - EX_i)}
\end{aligned} \tag{4}$$

Lemma 2.3. If random variable X has zero mean, $EX = 0$, and is bounded in [a,b], then for any $s > 0$,

$$Ee^{sX} \leq e^{s^2(b-a)^2/8}$$

Proof. By convexity of exponential function, we have

$$e^{sX} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-X}{b-a}e^{sa} \tag{5}$$

Taking expectation on (5), for $EX = 0$, we get

$$\begin{aligned} E(e^{sX}) \leq& \frac{be^{sa} - ae^{sb}}{b-a} \\ =& \big(1 - p + pe^{s(b-a)}\big)e^{-s(b-a)} \\ \equiv& e^{\phi(u)} \end{aligned} \tag{6}$$

Where $u = s(b-a)$. We know that $\phi(0) = \phi'(0) = 0$ and

$$\phi''(u) = \frac{p(1-p)e{-}u}{\big(p + (1-p)e^{-u}\big)^2} \leq \frac{1}{4}$$

So by Taylor expansion, for some $\theta \in [0, u]$

$$\phi(u) = \phi(0) + u\phi'(0) + u^2/2\phi''(\theta) \leq u^2/8 = \frac{s^2(b-a)^2}{8}$$

Hence,

$$Ee^{sX} \leq e^{s^2(b-a)^2/8}$$

□

Theorem 2.4. Hoeffding's Inequality

For bounded random variables $X_i \in [a_i, b_i]$, $X_1, \cdots, X_n$ are independent. Then

$$P\{S_n - ES_n \geq t\} \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

and

$$P\{S_n - ES_n \leq -t\} \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

Proof. Combine with (4) and Lemma 2.3

$$P\{S_n - ES_n \geq t\} \leq \exp(-st + \sum_{i=1}^n \frac{s^2(b_i - a_i)^2}{8}) \tag{7}$$

Set $s = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$ and minimize RHS in (7). That is

$$P\{S_n - ES_n \geq t\} \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

□

If we apply Hoeffding's inequality to a binomial distribution,

$$P(\hat{R}(f) - R(f) \geq \epsilon) = P(S_n/n - p \geq \epsilon) \leq \exp(-2\epsilon^2 n) \tag{8}$$

where $b_i - a_i = 1$. (8) decreases at almost the same rate as the central limit theorem. When $p = 1/2$, (8) is eaxactly what central limit theorem predicts. Meanwhile, in (8), we don't assume $R(f) \to 0$.

Theorem 2.5. Hoeffding/Azuma/McDiarmid's/Bounded difference inequality

Suppose random variables, $X_1, \cdots, X_n \in \mathcal{X}$, are independent. f is a mapping from X^n to $\mathbb{R}$. $\forall i, \forall x_1, \cdots, x_n, x_i' \in \mathcal{X}$, f satisfies

$$|f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)| \leq c_i \tag{9}$$

Then

$$P\{f(x_1, \cdots, x_n) - Ef \geq t\} \leq \exp(\frac{-2t^2}{\sum_{i=1}^n c_i^2})$$
$$P\{f(x_1, \cdots, x_n) - Ef \leq -t\} \leq \exp(\frac{-2t^2}{\sum_{i=1}^n c_i^2})$$

Proof. Define random variable $V_i = E(f|X_1, \cdots, X_i) - E(f|X_1, \cdots, X_{i-1})$. Then $f - Ef = \sum_{i=1}^n V_i$. Also, $E(V_i|X_1, \cdots, X_{i-1}) = 0$.

$$\begin{aligned} P(f - Ef \geq t) =& P(\sum_{i=1}^n V_i \geq t) \\ \leq& E\Big(e^{s\sum_{i=1}^n V_i} e^{-st}\Big) \\ =& e^{-st} E(\prod_{i=1}^n e^{sV_i}) \\ =& e^{-st} E(E(\prod_{i=1}^n e^{sV_i}|X_1, \cdots, X_{n-1})) \\ \cdots& \\ =& e^{-st} E(\prod_{i=1}^n E(e^{sV_i}|X_1, \cdots, X_{i-1})) \end{aligned} \tag{10}$$

Define

$$W_i = \sup_{X_i} \big(E(f|X_1, \cdots, X_i) - E(f|X_1, \cdots, X_{i-1})\big)$$
$$Z_i = \inf_{X_i} \big(E(f|X_1, \cdots, X_i) - E(f|X_1, \cdots, X_{i-1})\big)$$

Obviously, $W_i \leq V_i \leq Z_i$ and

$$W_i - Z_i = \sup_{X_i} \sup_{X_i'} (E(f|X_1, \cdots, X_i)) - E(f|X_1, \cdots, X_i')) \leq c_i$$

We know $E(V_i|X_1, \cdots, X_{i-1}) = 0$. Apply lemma 2.3 to (10), Hence,

$$P\{f - Ef \geq t\} \leq \inf_s \Big\{ \exp(-st + \frac{s^2}{8} \sum_{i=1}^n c_i^2) \Big\}$$

Set $S = 4t/\sum_{i=1}^n c_i^2$, we can obtain the final result:

$$P\{f(x_1, \cdots, x_n) - Ef \geq t\} \leq \exp(\frac{-2t^2}{\sum_{i=1}^n c_i^2})$$

Replacing f with $-f$, we get another inequality:

$$P\{f(x_1, \cdots, x_n) - Ef \leq -t\} \leq \exp(\frac{-2t^2}{\sum_{i=1}^n c_i^2})$$

□

Here are two applications of Hoeffding/Azuma/McDiarmid's/Bounded difference inequality.

1. Let $f(X_1, \cdots, X_n) = S_n = \sum_{i=1}^n X_i$, X_i are bounded random variables. $a_i \leq X_i \leq b_i$. Let $c_i = b_i - a_i$. Clearly, f satisfies the condition (9). And we can have the same inequality in (2.4).

2. Consider $f(x_1, \cdots, x_n) = \sup_{g \in F}(Eg - \hat{E}_n g)$, e.g. $g(x, y) = l(y, \hat{f}(x))$, Eg is true risk and $\hat{E}_n g$ is empirical risk. Suppose g is a bounded function. $g : X \to [a, b]$. Hence, the biggest change in f is bounded by $\frac{b-a}{n}$, when an arbitrary X_i varies in $[a, b]$, others are fixed. Let $c_i = \frac{b-a}{n}$.

Let $\delta = \exp\{\frac{-2t^2}{\sum_{i=1}^n c_i^2}\} = \exp\{\frac{-2nt^2}{(b-a)^2}\}$, or $t = (b-a)\sqrt{\frac{-\ln\delta}{2n}}$. By theorem 2.5, with probability at most δ,

$$f - Ef \geq t$$

That is

$$\sup_{g \in \mathcal{F}}(Eg - \hat{E}_n g) \leq E_{X_1, \cdots, X_n}\big(\sup_{g \in \mathcal{F}}(Eg - \hat{E}_n g)\big) + (b-a)\sqrt{\frac{-\ln\delta}{2n}}$$